

Action Recognition by Graph Embedding and Temporal Classifiers



Ehsan Zare Borzeshi

Faculty of Engineering and Information Technology

University of Technology, Sydney

A dissertation submitted for the degree of

Doctor of Philosophy

May 2014

Certificate of Authorship and Originality

Title: **Action Recognition by Graph Embedding and Temporal Classifiers**

Author: **Ehsan Zare Borzeshi**

Date: **May 1, 2014**

Degree: **PhD**

I certify that the work in this dissertation has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the dissertation has been written by me. Any help that I have received in my research work and the preparation of the dissertation itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the dissertation.

Signature of author

I would like to dedicate this dissertation to my beloved wife

Shima

for her unending love and support.

Acknowledgements

Many people have contributed in various ways to make this PhD study an exciting and memorable journey. Just to name a few:

I would like to acknowledge the lively and stimulating atmosphere in our group. The lunches, discussions at the coffee shops and evenings out: I have enjoyed all of them. Several persons deserve a special mention. I gratefully acknowledge all of my current and former group members: Dr. Oscar Perez Concha, Ava Bargi, Shaukat Abidi, Jaime Andres Garcia, Dr. Richard Yi Da Xu; my friends: Prof Federico Giroso and Dr. Majid Nazem; and members of the iNext research centre at UTS. Furthermore, I owe sincere gratitude to all members of the Center for Research in Computer Vision at UCF: Afshin Dehghan, Shayan Modiri and Amir Roshan Zamir; and Dr. Kasper Riesen at University of Applied Sciences and Arts Northwestern Switzerland.

I would like to express my utmost gratitude to my external advisors, Professor Mubarak Shah and Professor Horst Bunke for all their invaluable help and advice during my studies.

I would like to acknowledge ARC and my colleagues in the “Airports of the Future” project. This work has been supported by the Australian Research Council (ARC) under Linkage Projects Scheme “LP0990135”.

I would like to thank my dissertation reviewers for finding the time to read and comment on this work.

I would like to thank my parents, Mohammad and Marzieh, for giving birth to me at the first place, teaching me that it is no disgrace to work hard (which has proven to be a useful lesson) and supporting me spiritually throughout my life. I also thank them for their faith in me and allowing

me to be as ambitious as I wanted. It was under their watchful eye that I have gained so much drive and ability to tackle challenges head on. Furthermore, I thank my sister and brother, Elaheh and Amirhossein, for their love and support. I am so lucky to have all of you as my family.

I would also like to thank my wife's parents, Hossein and Tahereh, for first letting me take her hand in marriage, and for their extreme kindness and invaluable support to date. They are my best friends and beloved family and their interest and support has been tremendously valuable and appreciated indeed.

A huge thanks to my supervisor Professor Massimo Piccardi for supporting me during my PhD, for being patient and for being critical about my work. I admire his brilliance, determination, and hunger for knowledge. I have very much appreciated his pleasant way of supervising, even though I have never mentioned this explicitly. He has been an invaluable source of insights that any PhD student would love to know but which have never been written down. Massimo has supported me academically and emotionally through the rough road to finish this thesis. I feel deeply indebted to Massimo and a huge portion of the knowledge and confidence that I have today is due to him.

This list would be incomplete without expressing the most wholehearted gratitude to my life angel and love, Shima. Her support has been unconditional all these years; she has given up many things for me to finish my study; she has cherished with me every great moment and supported me whenever I needed it. Her positive spirit, unwavering love, patient endurance and tolerance of my occasional vulgar mood throughout the last ten years deserve so much more than just a "thank you". Not only now, but also in the years to come.

Abstract

With the improved accessibility to an exploding amount of video data and growing demand in a wide range of video analysis applications, video-based action recognition becomes an increasingly important task in computer vision. Unlike most approaches in the literature which rely on bag-of-feature methods that typically ignore the structural information in the data, in this monograph we incorporate the spatial relationship and the time stamps in the data in the recognition and classification processes.

We capture the spatial relationships in the subject performing the action by representing the actor's shape in each frame with a graph. This graph is then transformed into a vector of real numbers by means of prototype-based graph embedding. Finally, the temporal structure between these vectors is captured by means of sequential classifiers. The experimental results on a well-known action dataset (KTH) show that, although the proposed method does not achieve accuracy comparable to that of the best existing approaches, these embedded graphs are capable of describing the deformable human shape and its evolution over time.

We later propose an extended hidden Markov model, called the hidden Markov model for multiple, irregular observations (HMM-MIO), capable of fusing spatial information provided by graph embedding and the textual information of STIP descriptors. Experimental results show that recognition accuracy can be significantly improved by combining the spatio-temporal features with the structural information obtaining higher accuracy than from either separately. Furthermore, HMM-MIO is applied to the task of joint action segmentation and classification over a concatenated version of the KTH action dataset and the challenging CMU multi-modal activity dataset. The achieved accuracies proved comparable to or higher

than state-of-the-art approaches and show the usefulness of the proposed model also for this task.

The next and most remarkable contribution of this dissertation is the creation of a novel framework for selecting a set of prototypes from a labelled graph set taking class discrimination into account. Experimental results show that such a discriminative prototype selection framework can achieve superior results, not only for the task of human action recognition, but also in the classification of various structured data such as letters, digits, drawings, fingerprints compared to other well-established prototype selection approaches.

Lastly, we change our focus from the forementioned problems to the recognition of complex event, which is a recent area of computer vision expanding the traditional boundaries of visual recognition. For this task, we have employed the notion of concept as an alternative intermediate representation with the aim of improving event recognition. We model an event by a hidden conditional random field and we learn its parameters by a latent structural SVM approach. Experimental results over video clips from the challenging TRECVID MED 2011 and MED 2012 datasets show that the proposed approach achieves a significant improvement in average precision at a parity of features and concepts.

Contents

Contents	vii
List of Figures	xi
1 Introduction	1
1.1 Overview	1
1.2 Research Questions	2
1.3 Outline of the Dissertation	3
1.4 Publications	5
2 Literature Review	7
2.1 Action Recognition	7
2.1.1 Challenges	8
2.1.2 Feature Extraction	9
2.1.3 Action Detection and Classification	10
2.1.4 Joint Segmentation and Classification	11
2.2 Graph Theory	11
2.2.1 Graph	11
2.2.2 Graph Matching	12
2.2.2.1 Graph Edit Distance	13
2.2.2.2 Probabilistic Graph Edit Distance	14
2.2.2.3 Bipartite Graph Edit Distance	15
2.3 Graphical model based learning	16
2.3.1 Inference and Learning	16
2.3.2 Directed and Undirected Graphical Models	19

2.3.3	Probabilistic Graphical Models for Sequential Data	20
2.4	Support Vector Machine	24
3	Action Recognition by Graph Embedding	33
3.1	Prior Work and Our Contributions	33
3.2	Proposed Methods	34
3.2.1	Graph Embedding	35
3.2.2	Feature Extraction	35
3.2.3	Prototype Selection Techniques	38
3.2.4	Classification	41
3.3	Experimental Results	43
3.3.1	Evaluation of the feature vectors	43
3.3.2	Comparison to the state of the art	44
3.4	Discussion and Conclusions	46
4	Fusion of Texture and Structural Features for Action Recognition	48
4.1	Prior Work and Our Contributions	49
4.2	Proposed Methods	52
4.2.1	Classification and Time Segmentation	52
4.2.1.1	HMM-MIO	52
4.2.1.2	Scale of the observation probabilities in HMM-MIO	55
4.2.1.3	Forward and backward formulas for HMM-MIO . .	56
4.2.1.4	A brief comparison with discriminative sequential models	57
4.2.1.5	Experimental Results	58
4.2.2	Feature Fusion	62
4.2.2.1	Features	62
4.2.2.2	Fusion graphical model	63
4.2.2.3	Experimental Results	63
4.3	Discussion and Conclusions	64
5	Discriminative Prototype Selection	66
5.1	Prior Work and Our Contributions	66
5.2	Proposed Methods	69

CONTENTS

5.2.1	Prototype selection	69
5.2.2	Learning discriminative prototypes	70
5.2.3	Discriminative prototype selection algorithms	70
5.2.3.1	Discriminative Center Prototype Selection	71
5.2.3.2	Discriminative Border Prototype Selection	71
5.2.3.3	Discriminative Repelling Prototype Selection	72
5.2.3.4	Discriminative Spanning Prototype Selection	73
5.2.3.5	Discriminative Targetsphere Prototype Selection	73
5.2.3.6	Discriminative k -Center Prototype Selection	74
5.3	Experimental Results	74
5.3.1	Dataset	74
5.3.1.1	Letter datasets	75
5.3.1.2	Digit dataset	76
5.3.1.3	GREC dataset	76
5.3.1.4	Fingerprint dataset	77
5.3.1.5	AIDS data set	78
5.3.1.6	Mutagenicity dataset	78
5.3.1.7	Protein dataset	79
5.3.1.8	Webpage dataset	80
5.3.2	Comparison between the discriminative and labeled approaches	80
5.4	Discussion and Conclusions	86
6	Complex Event Recognition by Latent Temporal Models of Concepts	88
6.1	Prior Work and Our Contributions	89
6.2	Proposed Methods	92
6.2.1	Latent State Initialization	95
6.2.2	Time-Sparsity of Concepts	95
6.3	Experimental Results	97
6.3.1	TRECVID MED 2011 Event Collection	101
6.3.2	TRECVID MED 2012 Event Collection	103
6.4	Discussion and Conclusions	106
7	Conclusions	107

CONTENTS

References	110
------------	-----

List of Figures

1.1	A visual sketch of the approaches presented in the various chapters.	4
2.1	An example edit path between g_1 and g_2 (node labels are represented by different shades of gray). Image courtesy of Kaspar Riesen [111].	13
2.2	A simple directed graphical model	18
2.3	Graphical model for the HMM	21
2.4	Graphical model for Classification with HMM	22
2.5	Graphical model for the Linear-chain CRF	23
2.6	Graphical model for the HCRF	23
2.7	Example of SVM classification (linear separable case)	25
2.8	A single outlier point can significantly affect the separating hyperplane significantly	26
2.9	Non-separable classes	27
2.10	Structured-output SVMs	29
2.11	Multi-class SVMs	30
2.12	Latent structured-output SVMs	31
3.1	KTH human action database: examples of sequences corresponding to different types of actions and scenario [124].	36
3.2	Bounding box generated from a modified tracker [25] using the KTH action dataset and the extracted SIFT keypoints composed into a graph.	37
3.3	Examples of selected postures from the KTH action dataset.	38
3.4	The time-sequential values of a 19-dimensional feature vector obtained from graph embedding based on the $c - dps$ for one action (boxing) performed by one subject in the KTH action dataset.	39

3.5	Illustration of the different prototype selectors applied to the training set. The number of prototypes is defined by $N = 30$. The prototypes selected by the respective selection algorithms are shown with red dots. Image courtesy of Kaspar Riesen [111].	40
3.6	Instance images illustrate that the SIFT keypoints are not able to capture the body shape sufficiently well to be used as a shape descriptor. .	47
4.1	Example of the spatio-temporal interest points from [72] in a video from the KTH action dataset. Frames are displayed in row-major order. The radius of circles is proportional to the scale at which change is detected. Note the variable number of points appearing in subsequent frames.	50
4.2	(a) Decoding the state sequence, $y_{1:T}$, of an HMM provides joint action classification and segmentation from observations $x_{1:T}$; (b) decoding variable a by Bayes' inversion rule and marginalization of $y_{1:T}$ provides a single action label for the entire sequence $x_{1:T}$	53
4.3	The uniform grid over the actor's area.	54
4.4	The generative model of HMM-MIO.	55
4.5	Examples of actions for preparation of "brownies": (from left to right, column wise) <i>close</i> , <i>crack</i> , <i>none</i> , <i>open</i> , <i>pour</i> , <i>put</i> , <i>read</i> , <i>spray</i> , <i>stir</i> , <i>switch-on</i> , <i>take</i> , <i>twist-off</i> , <i>twist-on</i> and <i>walk</i>	62
4.6	Modified HMM-MIO (hidden Markov model with multiple, independent observations); x_t are the observations at time t (appearance observations provided by the STIP descriptors, x_a , and the structural observation provided by graph embedding, x_s); y_t is the corresponding hidden state; W_a and W_s are the two weights for computing the total observation probability $P(x_t y_t) = W_a \cdot P_a(x_{a,t}^{1:N_t} y_t) + W_s \cdot P(x_{s,t} y_t)$; $W_a + W_s = 1$	64
5.1	Examples of letter A: Original and distortion levels low, medium and high (from left to right)	76
5.2	A graph example of each of the ten digit classes	76
5.3	An instance image of each distortion level	77

LIST OF FIGURES

5.4	Instances of fingerprint classes: left, right, arch and whorl (from left to right)	78
5.5	A molecular compound of both classes: active and inactive (from left to right)	79
5.6	An example of each class: EC1, EC2, EC3, EC4, EC5 and EC6 (from left to right)	80
5.7	Accuracy with various prototype selection approaches and datasets as a function of the number of prototypes per class. (a) Letter High, l-sps vs d-sps; (b) Digit, l-sps vs d-sps; (c) Grec, l-cps vs d-cps; (d) Letter Medium, l-bps vs d-bps; (e) Letter Low, l-tps vs d-tps; (f) Mutagenicity, l-sps vs d-sps.	84
5.8	Accuracy with various prototype selection approaches and datasets as a function of the value of W_s (The reported W_s value is multiplied by 100). (a) Letter Medium, d-bps; (b) Mutagenicity, d-sps;.	85
6.1	A birthday party event and its articulation over concepts.	90
6.2	The graphical model of the hidden conditional random field. Variable a is the event class, $y_{1:T}$ are the latent states and $x_{1:T}$ are the measurements (output of concept detectors in this work).	93
6.3	Time-sparsity of concepts and states. The top plot shows the output of concept detectors above 0.4 for an event of type “Dog show”. The bottom plot shows the corresponding trellis of the states. Sparsity is evident in both the concept detectors’ outputs and the states.	96
6.4	Examples from complex video event categories.	98